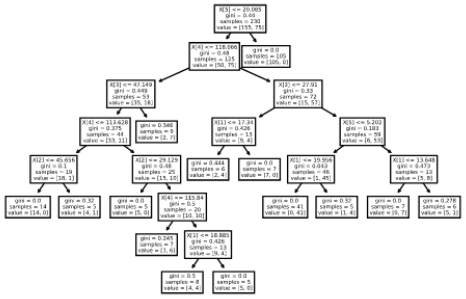
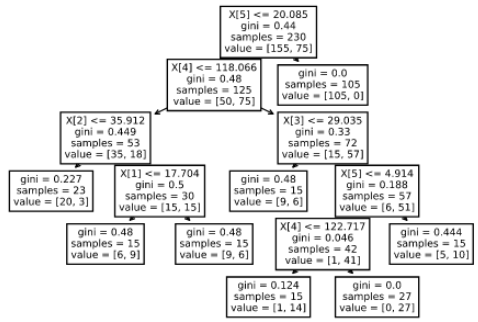
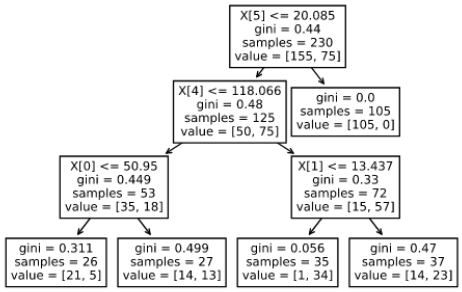
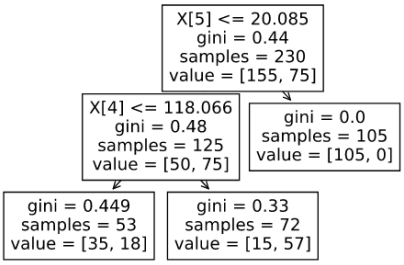
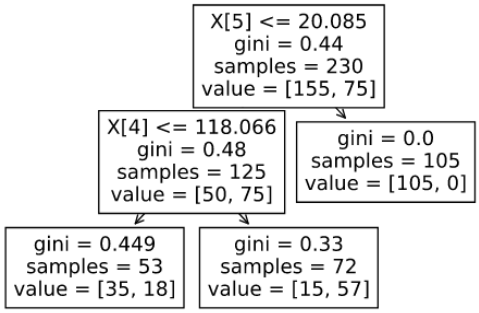
1. (a) The decision trees were constructed by DecisionTreeClassifier from Scikit.
The five trees are shown below, ordered from left to right, and top to bottom.



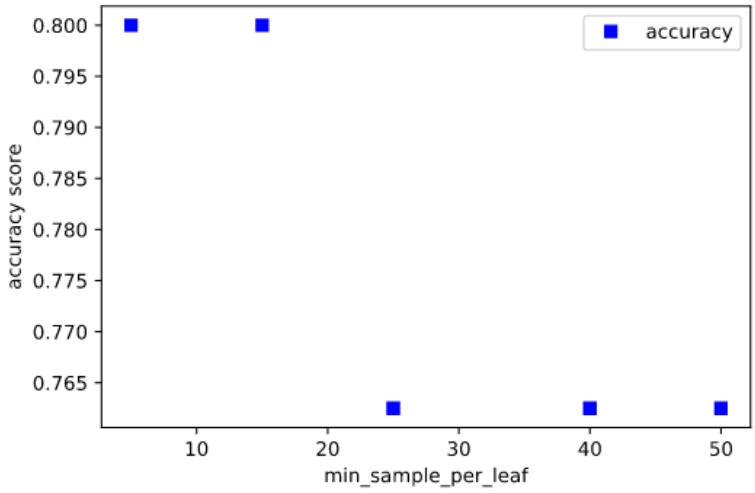








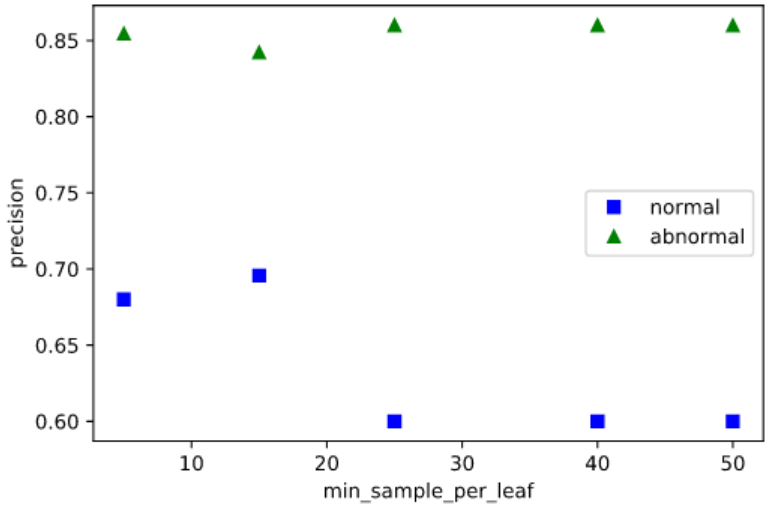
The accuracy scores are shown below.



The value of min_samples_leaf affects the depth of the tree. Since “a split point at any depth will only be considered if it leaves at least min_samples_leaf training samples in

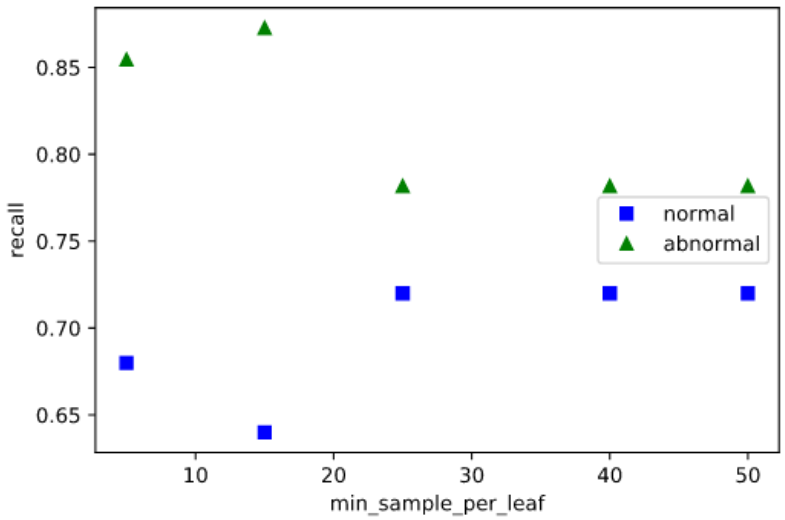
each of the left and right branches". The min_samples_leaf = 15 is the best choice among all the options. The min = 5 might has over-fitting. The min = 25, 40 and 50 are under-fitting since not all features are used in splitting.

(b) The precision values are plotted below.



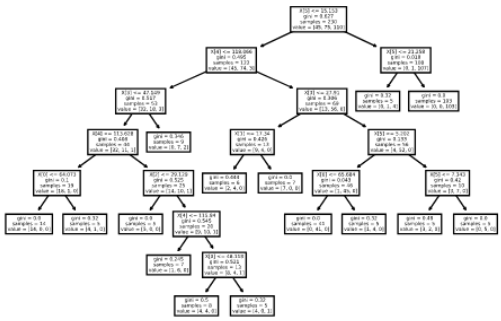
The abnormal precision is less affected by the under fitting than the normal precision. The reason could be the features selected for splitting split the abnormal class better than the normal case.

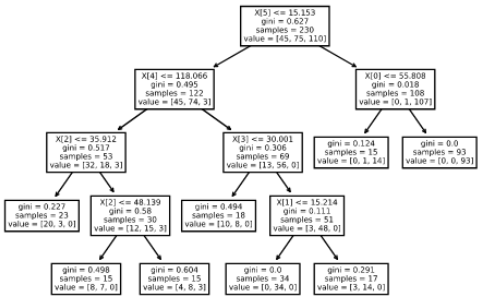
The recall values are plotted below.

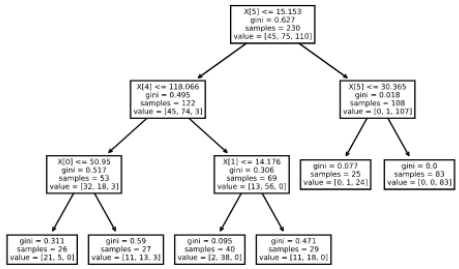


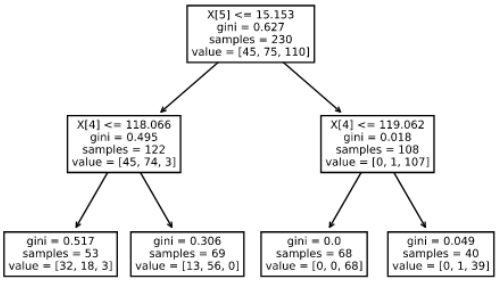
The precision/recall for normal and abnormal have opposite trend with respect to min_sample_per_leaf.

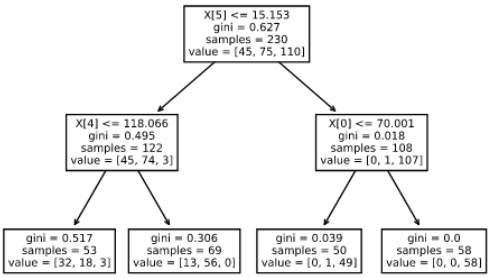
2. (a)

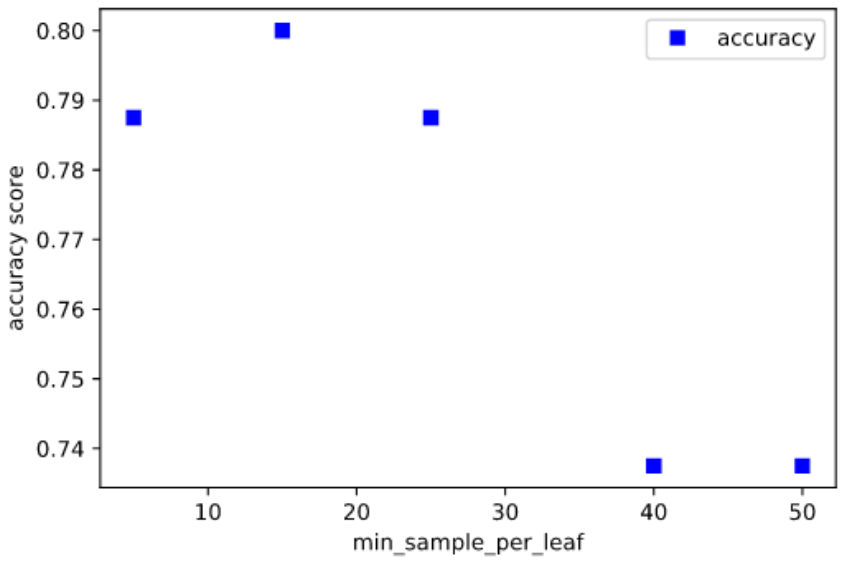



X[5] <= 15.153
gini = 0.627
samples = 230
value = [45, 75, 110]
X[4] <= 118.066
gini = 0.495
samples = 122
value = [45, 74, 3]
X[0] <= 55.808
gini = 0.018
samples = 108
value = [0, 1, 107]
X[2] <= 35.912
gini = 0.517
samples = 53
value = [32, 18, 3]
X[3] <= 30.001
gini = 0.306
samples = 69
value = [13, 56, 0]
gini = 0.124
samples = 15
value = [0, 1, 14]
gini = 0.0
samples = 93
value = [0, 0, 93]
gini = 0.227
samples = 23
value = [20, 3, 0]
X[2] <= 48.139
gini = 0.58
samples = 30
value = [12, 15, 3]
gini = 0.494
samples = 18
value = [10, 8, 0]
X[1] <= 15.214
gini = 0.111
samples = 51
value = [3, 48, 0]
gini = 0.498
samples = 15
value = [8, 7, 0]
gini = 0.604
samples = 15
value = [4, 8, 3]
gini = 0.0
samples = 34
value = [0, 34, 0]
gini = 0.291
samples = 17
value = [3, 14, 0]


X[5] <= 15.153
gini = 0.627
samples = 230
value = [45, 75, 110]
X[4] <= 118.066
gini = 0.495
samples = 122
value = [45, 74, 3]
X[5] <= 30.365
gini = 0.018
samples = 108
value = [0, 1, 107]
X[0] <= 50.95
gini = 0.517
samples = 53
value = [32, 18, 3]
X[1] <= 14.176
gini = 0.306
samples = 69
value = [13, 56, 0]
gini = 0.077
samples = 25
value = [0, 1, 24]
gini = 0.0
samples = 83
value = [0, 0, 83]
gini = 0.311
samples = 26
value = [21, 5, 0]
gini = 0.59
samples = 27
value = [11, 13, 3]
gini = 0.095
samples = 40
value = [2, 38, 0]
gini = 0.471
samples = 29
value = [11, 18, 0]

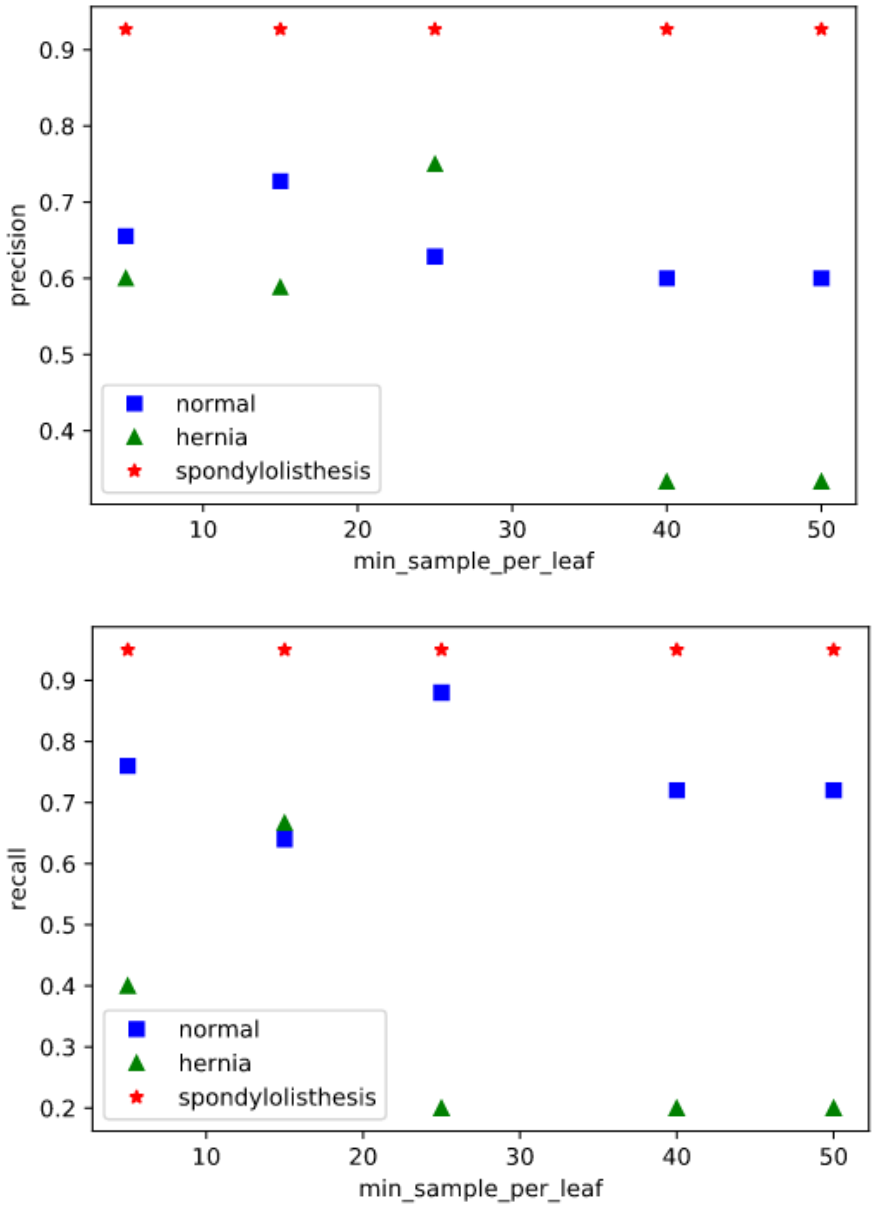

X[5] <= 15.153
gini = 0.627
samples = 230
value = [45, 75, 110]
X[4] <= 118.066
gini = 0.495
samples = 122
value = [45, 74, 3]
X[4] <= 119.062
gini = 0.018
samples = 108
value = [0, 1, 107]
gini = 0.517
samples = 53
value = [32, 18, 3]
gini = 0.306
samples = 69
value = [13, 56, 0]
gini = 0.0
samples = 68
value = [0, 0, 68]
gini = 0.049
samples = 40
value = [0, 1, 39]


X[5] <= 15.153
gini = 0.627
samples = 230
value = [45, 75, 110]
X[4] <= 118.066
gini = 0.495
samples = 122
value = [45, 74, 3]
X[0] <= 70.001
gini = 0.018
samples = 108
value = [0, 1, 107]
gini = 0.517
samples = 53
value = [32, 18, 3]
gini = 0.306
samples = 69
value = [13, 56, 0]
gini = 0.039
samples = 50
value = [0, 1, 49]
gini = 0.0
samples = 58
value = [0, 0, 58]


accuracy
0.80
0.79
0.78
0.77
0.76
0.75
0.74
accuracy score
10
20
30
40
50
min_sample_per_leaf

The best choice is 15. The depth of trees are similar to the 2-class trees. The over fitting is more significant than 2-class tree. However, the under fitting occurs later than that of the 2-class tree due to the increased classes. (The 3-class tree classifies the data better than 2-class tree for the 25 case.)

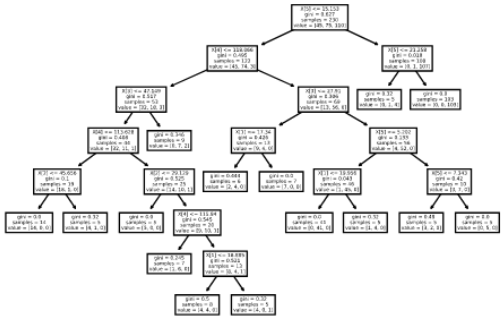
(b)

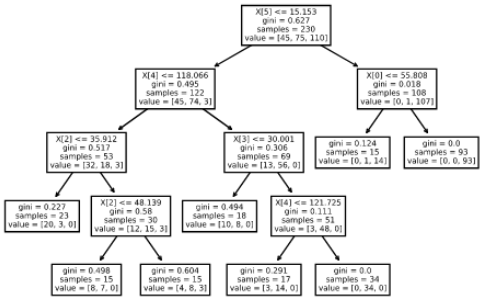


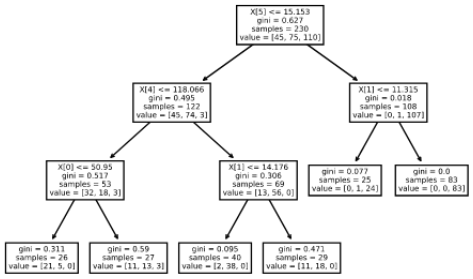
The Spondylolisthesis is easy to classify. The Hernia only correlates to certain features because it is sensitive to under fitting. The precision of normal is similar to the 2-class tree. The recall of normal is slightly higher than the 2-class tree. (False negative is reduced.)

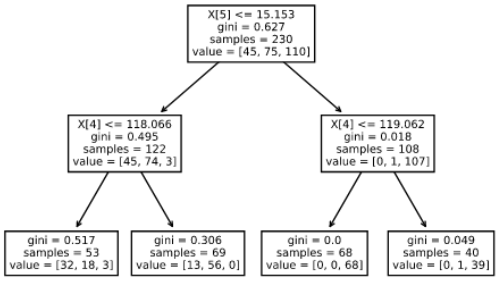
3. (a)The correlation matrix is shown below. The “degree_spond” has the highest correlation values. So it is dropped out of the data set.

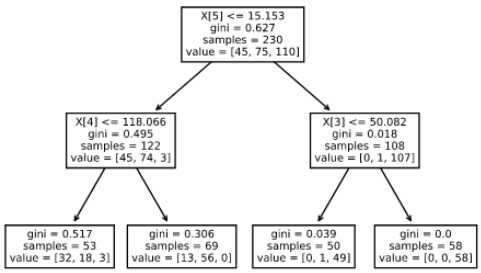
	pelvic_incidence	pelvic_tilt numeric	lumbar_lordosis_angle	sacral_slope	pelvic_radius	degree_spondylolisthesis	class_n
pelvic_incidence	1.000000	0.629199	0.717282	0.814960	-0.247467	0.638743	0.353336
pelvic_tilt numeric	0.629199	1.000000	0.432764	0.062345	0.032668	0.397862	0.326063
lumbar_lordosis_angle	0.717282	0.432764	1.000000	0.598387	-0.080344	0.533667	0.312484
sacral_slope	0.814960	0.062345	0.598387	1.000000	-0.342128	0.523557	0.210602
pelvic_radius	-0.247467	0.032668	-0.080344	-0.342128	1.000000	-0.026065	-0.309857
degree_spondylolisthesis	0.638743	0.397862	0.533667	0.523557	-0.026065	1.000000	0.443687
class_n	0.353336	0.326063	0.312484	0.210602	-0.309857	0.443687	1.000000

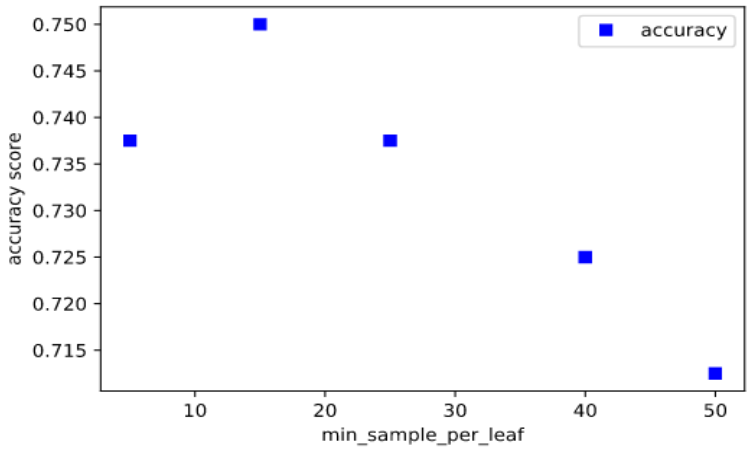






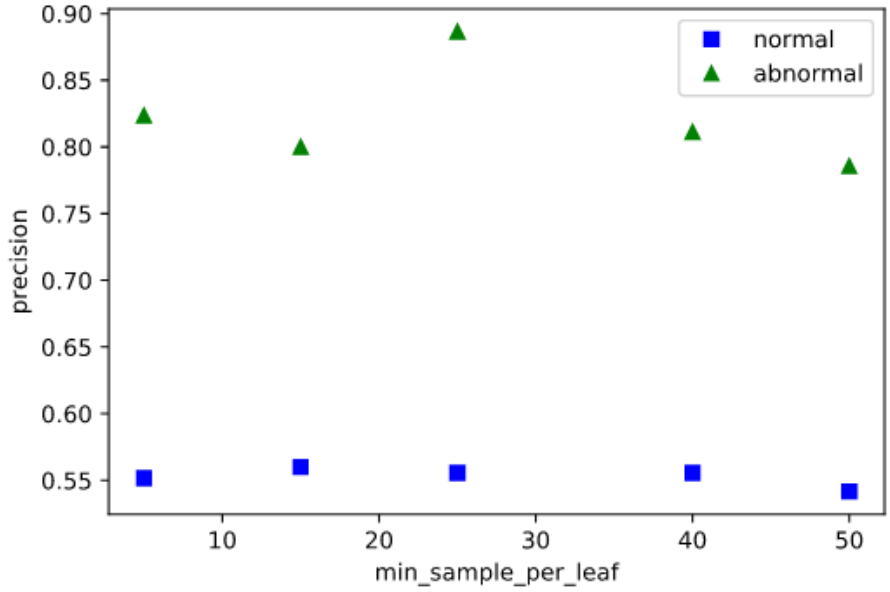


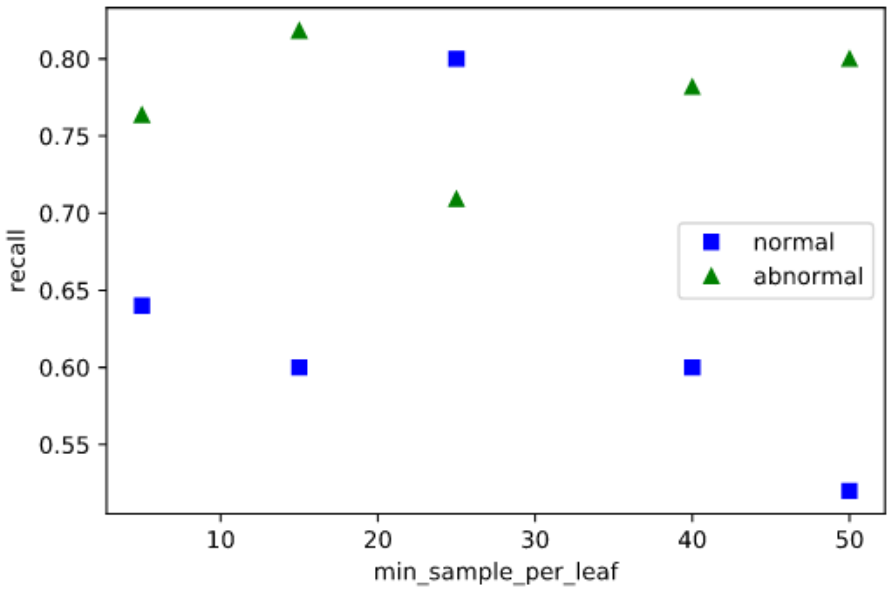




The best choice is still min = 15. Over fitting for min = 5 and under fitting for min = 25 and above.

(b)





The abnormal has a higher classification accuracy, and it is less affected by the over/under fitting. For the recall, normal and abnormal has opposite trend.

(c) The drop out simplified the tree at the cost of accuracy.

The drop out reduces the depth of the tree.

The precision accuracy is lower in 3 compared with 1. The best case in 3 is about the same as the under fitting case in 1.

The overall precision and recall are lower than that of the case in (1b). However, the trend with respect to min_per_leaf is similar.